Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., Angedakin, S., Casals, F., Navarro, A., Vigilant, L., Kühl, H. S., Langergraber, K., Boesch, C., Hughes, D., & Marques-Bonet, T. (2018). The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Molecular Ecology Resources*, *18*(2), 319-333. https://doi.org/10.1111/1755-0998.12728

Received: 29 June 2017 | Revised: 6 October 2017 | Accepted: 16 October 2017

DOI: 10.1111/1755-0998.12728

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

The impact of endogenous content, replicates and pooling on genome capture from faecal samples

Jessica Hernandez-Rodriguez[1] | Mimi Arandjelovic[2] | Jack Lester[2] | Cesare de Filippo[2] | Antje Weihmann[3] | Matthias Meyer[3] | Samuel Angedakin[2] | Ferran Casals[4] | Arcadi Navarro[1,5,6] | Linda Vigilant[2] | Hjalmar S. Kühl[2,7] | Kevin Langergraber[8] | Christophe Boesch[2] | David Hughes[1,9] | Tomas Marques-Bonet[1,5,6]

[1]Departament de Ciencies Experimentals i de la Salut, Institut de Biologia Evolutiva (Universitat Pompeu Fabra/CSIC), Barcelona, Spain

[2]Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[3]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[4]Genomics Core Facility, Departament de Ciencies Experimentals i de la Salut, Universitat Pompeu Fabra, Parc de Recerca Biomèdica de Barcelona, Barcelona, Spain

[5]Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain

[6]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[7]German Centre for Integrative Biodiversity Research (iDiv) Halle-Leipzig-Jena, Leipzig, Germany

[8]School of Human Evolution & Social Change, Arizona State University, Tempe, AZ, USA

[9]MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

Correspondence
David Hughes
Email: hughes.evoanth@gmail.com
and Tomas Marques-Bonet,
Email: tomas.marques@upf.edu

Funding information
NIH Clinical Center, Grant/Award Number: MH106874; Ministerio de Economía y Competitividad (MINECO) Spain, Grant/Award Number: BFU2015-68649-P, BFU2014-55090-P; Fondo Europeo de Desarollo Regional (FEDER) Spain, Grant/

Abstract

Target-capture approach has improved over the past years, proving to be very efficient tool for selectively sequencing genetic regions of interest. These methods have also allowed the use of noninvasive samples such as faeces (characterized by their low quantity and quality of endogenous DNA) to be used in conservation genomic, evolution and population genetic studies. Here we aim to test different protocols and strategies for exome capture using the Roche SeqCap EZ Developer kit (57.5 Mb). First, we captured a complex pool of DNA libraries. Second, we assessed the influence of using more than one faecal sample, extract and/or library from the same individual, to evaluate its effect on the molecular complexity of the experiment. We validated our experiments with 18 chimpanzee faecal samples collected from two field sites as a part of the Pan African Programme: The Cultured Chimpanzee. Those two field sites are in Kibale National Park, Uganda ($N = 9$) and Loango National Park, Gabon ($N = 9$). We demonstrate that at least 16 libraries can be pooled, target enriched through hybridization, and sequenced allowing for the genotyping of 951,949 exome markers for population genetic analyses. Further, we observe that molecule richness, and thus, data acquisition, increase when using multiple libraries from the same extract or multiple extracts from the same sample. Finally, repeated captures significantly decrease the proportion of off-target reads from 34.15% after one capture round to 7.83% after two capture rounds, supporting our conclusion that two rounds of target enrichment are advisable when using complex faecal samples.

KEYWORDS
conservation genetics, exome, next-generation sequencing, noninvasive samples, target enrichment

Award Number: SAF2012-35025, SAF2015-68472-C2-2-R; Howard Hughes Medical Institute; International Early Career; Max Planck Society Innovation Fund; Heinz L. Krekeler Foundation's; Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya

1 | INTRODUCTION

Over the last few years, there has been a growing interest in the use of noninvasive (NI) samples such as hair and faeces for studying the population genomics of wild animal populations (Ouborg, Pertoldi, Loeschcke, Bijlsma, & Hedrick, 2010; Primmer, 2009; Shafer et al., 2015; Steiner, Putnam, Hoeck, & Ryder, 2013). The use of NI samples is preferable for understanding animal population histories for two main reasons. First, by noninvasively collecting samples, no physical harm comes to the animal. This is in contrast to attempts of collecting blood or other tissues which also increases the risk of infection, elevates an individuals' stress and can alter behaviour and social group dynamics (Morin, Wallis, Moore, Chakraborty, & Woodruff, 1993; Taberlet, Waits, & Luikart, 1999). Second, the ability to use NI samples limits the need to rely upon samples collected from zoos, museums, sanctuaries or hunted animals. While such samples remain vital for a variety of research efforts, they are not always ideal, often lacking information on the geographic origin of the sample, and do not necessarily represent extant diversity of the species (Hofreiter, Siedel, Van Neer, & Vigilant, 2003; Yu, Jensen-Seaman, Chemnick, Ryder, & Li, 2004). The two major disadvantages of NI samples are their (1) low endogenous DNA content and (2) their degraded DNA (Perry, Marioni, Melsted, & Gilad, 2010). NI samples are generally a composite of genetic material derived from an individuals' own cells and from microorganisms living within, on, and/or around the biological source material, acting as a substrate for the nonendogenous DNA contributors. Further, NI samples are often collected in warm, humid environments that negatively impact the quality of cellular material over time. Resultantly, NI samples are not the most ideal source material for acquiring endogenous nucleic acids. For these reasons, studies using NI samples have been restricted to targeting a limited number of markers or genetic loci. Nevertheless, population genetic studies in great apes have been vitally successful in genotyping autosomal microsatellites (Fünfstück et al., 2014, 2015; Inoue et al., 2013; Kanthaswamy, Kurushima, & Smith, 2006; Morin et al., 1993; Nater et al., 2013; Thalmann, Fischer, Lankester, Pääbo, & Vigilant, 2007), Y-chromosome microsatellites (Arandjelovic et al., 2011; Eriksson et al., 2006; Erler, Stoneking, & Kayser, 2004; Langergraber et al., 2014), autosomal regions (Fischer, Wiebe, Pääbo, & Przeworski, 2004; Fischer et al., 2011; Hans et al., 2015; Thalmann et al., 2007) and the high copy number mitochondrial genome (Thalmann, Hebler, Poinar, Pääbo, & Vigilant, 2004; Thalmann, Serre, et al., 2004) from NI samples. These PCR-based targeted genetic efforts are not limited to anthropologist but are common to all biologists in a variety of subdisciplines (Swenson, Taberlet, & Bellemain, 2011; Wultsch, Waits, Hallerman, & Kelly, 2015; Wultsch, Waits, & Kelly, 2014), all of which could be aided by new techniques that could provide more data. To date, blood and other tissue sources have been widely used in genetic and population history studies (Lobon et al., 2016; de Manuel et al., 2016; Prado-Martinez et al., 2013; Rogers & Gibbs, 2014; Xue et al., 2016), and given the quality and quantity of such DNA, they will maintain their vital role in molecular research. However, in this current study, we take another step towards attenuating our dependency upon such samples for acquiring deep genomic data and improve upon the ability of biologist to study the genetic diversity of wild, extant populations, while minimizing direct interaction and contact.

Recent target enrichment methodologies have provided methodological advances in acquiring more information from NI samples (Perry et al., 2010; Snyder-Mackler et al., 2016; Wall et al., 2016). These enrichment methods are performed with the use of biotinylated RNA baits that hybridize with the DNA from species of interest, which are subsequently isolated and sequenced. These studies exemplify the potential of these methodologies for evolutionary, ecological, population and conservation genetic efforts.

Consequently, we used a commercial kit from Roche to target-capture enrich and sequence the chimpanzee exome (57.5 Mb). The study design was chosen to allow for: (1) an evaluation of multiplex hybridization enrichment; (2) comparison between one and two rounds of hybridization enrichment; (3) the quantification of sample quality, defined here as the endogenous DNA content and level of DNA fragmentation, on performance; (4) measuring discordance among (a) hybridization replicates, (b) library replicates, (c) extract replicates and (d) faeces replicates; and finally (5) evaluating the potential utility of using replicates to increase data output. We have chosen to target the exome as it represents, relative to the genome, a small target space, which in this study is at 57.5 Mb. Moreover, with it being the protein-coding portion of the genome, it is a prime target space for studies of natural selection, protein function and evolution and yet, also remains useful in estimations of population ancestry, inbreeding and potential geographic assignment. To the best of our knowledge, this study design is the first to explicitly evaluate the performance and difficulties of pooling multiple, complex NI samples for target enrichment of an exome, while also having been simplified by the utilization of a commercial kit. The goal of all these experiments is to provide a knowledge base and some basic guidelines and recommendations for biologists in the use of NI samples in their own genetic studies.

2 | MATERIALS AND METHODS

2.1 | Samples

This study employed 18 faecal samples derived from 17 individuals previously collected as a part of the Pan African Programme: The Cultured Chimpanzee project (PanAf; http://panafrican.eva.mpg.de; Kühl et al., 2016; Vaidyanathan, 2011). All PanAf chimpanzee faecal samples are collected from unhabituated chimpanzees from up to 3-day-old faecal piles using a two-step ethanol-silica preservation method (Nsubuga et al., 2004). An initial subset of 48 collected samples was chosen as an initial screening panel. These 48 samples were chosen because they had previously performed well in other microsatellite genotyping assays indicating that they contained little to no inhibitory molecules (Arandjelovic et al., 2009, 2011). This was a minimal standard taken here to identify those samples of reasonable quality that should present no problems during library production. Arguably, a necessary step to limit the influence of inhibitors of PCR that may also detrimentally influence library preparation.

Each sample of the screening pool then had its' level of DNA degradation and endogenous DNA content measured. Here, degradation is the length distribution of DNA molecules, specifically we focused on the mean observed fragment length, and endogenous DNA content is defined as DNA derived from the source individuals' cells as opposed to gut microbial flora and/or environmental contaminants. Degradation was evaluated by running samples on a Fragment analyzer™ (Automated CE System 96 capillary, Advanced Analytical Technologies, Inc.), an automated system for the quantification and qualification of next-generation sequencing (NGS) libraries, genomic DNA (gDNA) and RNA, following the manufacturer's instructions for the High Sensitivity Genomic DNA Analysis Kit (Cat. Number DNF-488). Endogenous content was estimated by both qPCR and low-depth shotgun sequencing of sample libraries. Libraries for low-depth shotgun sequencing were prepared, for each sample, using published protocols for in-house library preparation (Meyer & Kircher, 2010).

From the screening pool, we chose 18 samples spanning the range of observed average fragmentation length and percentage of endogenous content. Samples were selected to span the range of these two quality summary statistics (Appendix S1: Fig. S1). Two of these 18 faecal samples are derived from a single chimpanzee individual (Figure 1, Exp.2), as determined by microsatellite genotyping carried out in independent unpublished work prior to this study (K. E. Langergraber, unpublished data). From each of these two faecal samples, we performed a second DNA extraction for the purpose of our second experimental design, outlined below. Neither of these two new DNA extracts were processed through the fragment analyser nor the endogenous content evaluated by low-depth shotgun sequencing. In total, these 18 faecal samples resulted in a total of 20 faecal DNA (fDNA) extracts (Figure 1) representing 17 unique individuals.

All 18 faecal samples are derived from collections carried out at two different locations. Nine samples of the *Pan troglodytes troglodytes* subspecies were collected from Loango National park, Gabon (Arandjelovic et al., 2011) and nine unpublished samples of the *Pan troglodytes schweinfurthii* subspecies from Kibale National Park, Uganda (K. E. Langergraber, unpublished data).

2.2 | Experimental designs

2.2.1 | Experiment 1

To assess the performance and replicability of a capture enrichment experiment involving a pool of multiple individuals, we followed the subsequent steps. Sixteen individually indexed libraries, each deriving from a unique individual, were collected into a single master pool at an equimolar ratio. That pool was then split into three equal pools or "replicates." Each replicate pool then went through two rounds of target enrichment prior to sequencing on a lane of the HiSeq 2500 separate from the others. In the end, this experiment yielded data for 48 experimental units (Figure 1).

2.2.2 | Experiment 2

The second experimental design was crafted to quantify the impact of wet laboratory technical variation on data acquisition and genotype discordance of a single sample, but also (1) to directly compare the realization of a single capture to that of a double capture and (2) to explore the information that may be gained by having faecal replicates, extract replicates, and/or library replicates in a study design. We define a faecal replicate as two or more unique faeces derived from a single chimpanzee. Extract replicates are two or more DNA extractions from a single faecal sample, and library replicates are two or more libraries produced from a single DNA extraction. Starting from a single chimpanzee, we identified two faecal samples derived from it (faecal or sample replicates). Then, from each of the sample replicates, we produced two DNA extracts (extract replicates), and from each extract, we produced two libraries (library replicates). Thus, from a single individual, we have a total of two faecal samples, four DNA extractions and eight uniquely indexed fDNA libraries (Figure 1). From these eight libraries, we made two equimolar library pools. Library replicates derived from a single extract went into different pools. This was carried out to ensure that any one replicate level (library, extract or sample) was not correlated with the downstream enrichment experiment. Finally, each of the two pools was then subdivided into three equal pools. To evaluate the execution of multiple rounds of capture the first pool was captured once, the second and third pools were captured twice. In the end, this experiment yielded data for 24 experimental units (Figure 1), and across both experiments, we total 72 experimental units.

2.3 | Library preparation, hybridization and capture

Compared to what would be expected from DNA isolated from a fresh tissue source, all samples presented degradation. Across the initial 48 samples present in the screening panel, we observe a broad

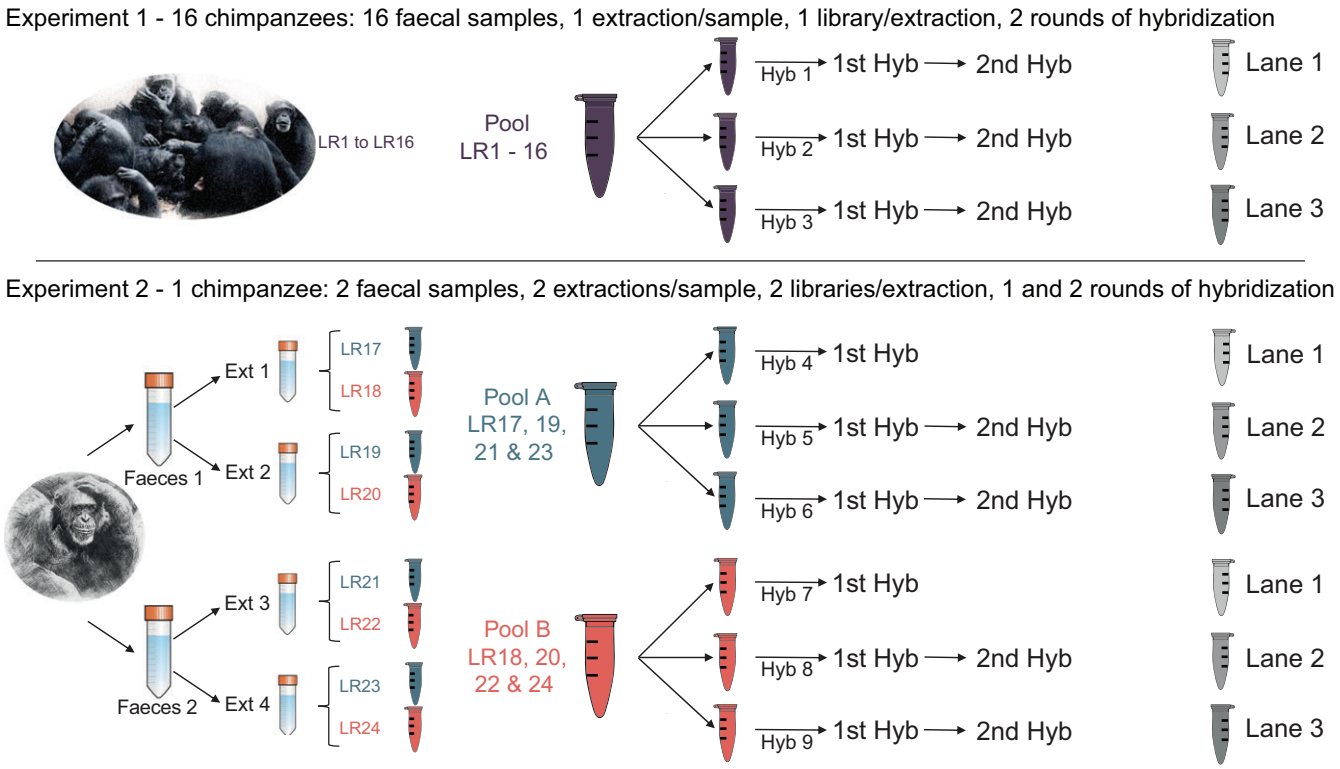

FIGURE 1 Experiments scheme. Experiment 1: Sixteen libraries were pooled and captured in triplicate. Experiment 2: Two faecal samples from a single chimpanzee, each extracted twice resulting in four extracts. Two libraries from each of the extracts were created resulting in eight libraries. Libraries were combined into two pools (A, B) such that each extract is present once in each pool. Each pool then underwent a single round and two double rounds of capture

range of fragment lengths, with large inter- and intrasample variability. Average DNA fragment lengths across samples ranged from 23 to 3,700 bp. Nevertheless, for library preparation, each sample required further shearing to acquire a more normally distributed samples with average lengths around 200 bp (Appendix S1: Fig. S2). Genomic DNA was fragmented by shearing using a Covaris S2 focused ultrasonicator with the following settings for 200-bp fragments: intensity 5, duty cycle 10%, cycles per burst 200, treatment time 120 s, temperature 7°C and water level 12. One concern about the fragmentation was that highly degraded samples, that is, those with lengths already near or below the target size of 200 bp, would be broken to shorter pieces. However, we did not observe this to be the case with the Bioanalyzer (Appendix S1: Fig. S2), and in addition, Covaris Inc. demonstrates that the smaller the size of the input DNA, an exponentially higher amount of energy is required to fragment it (http://covarisinc.com/resources/faqs/).

We built libraries, or indexed catalogs of the DNA molecules of our NI samples bound by known DNA sequences that allow for their sequencing, using the KAPA Library Preparation Kits (Cat. Number 07137923001). The SeqCap EZ Library SR User's Guide Version 5.1 was followed with some modifications described below. (1) Amount of starting material: for *Experiment 1*, we took 40 μl of each sample extract and added elution buffer up to a total volume of 53 μl (total DNA amount varied between 0.36 and 4.46 μg); for *Experiment 2*, we took 20 μl of each sample and added elution buffer up to a total of 53 μl (total DNA amount varied between 1.31 and 4.03 μg). (2) Reaction clean-ups for the end-repair and A-tailing were performed with MinElute Reaction clean-up spin columns (Cat. Number 28206) rather than Agencourt AMPure XP beads. This choice was made in an attempt to retain molecules smaller than 100 bp that could be overly abundant because of initial sample degradation. In our hands, the MinElute Reaction kit retains molecules down to ~50 bp, while SPRI-beads retained molecules down to ~100 bp. After each clean-up step, DNA was eluted in 20 μl of elution buffer. Reaction clean-ups in columns may be advantageous for those who chose to modify or even not perform the size selection step detailed in the commercial user guide, because of excessive sample degradation. With appreciable variation in degradation patterns across samples and for simplicity and comprehensiveness evaluation of this protocol, we have chosen to perform the size selection subsequent to ligation clean-up. After the ligation reaction, the first bead clean-up was performed using 90 μl of Agencourt AMPure XP beads, and the following steps were performed following the protocol. (3) Amplification of each sample library was performed using the precapture LM-PCR program, with a total of 12 cycles.

Libraries were quantified with an Agilent 2100 Bioanalyzer and DNA 1000 Assay kits and finally pooled depending on the experiment. *Experiment 1:* Equimolar pool of the sixteen libraries;

Experiment 2: Two pools each containing four libraries, one from each extract (Figure 1).

Nimblegen baits (Roche) for the chimpanzee exome (57.5 Mb) were designed using the panTro4 assembly (SeqCap EZ Developer Library, Cat Number 06740278001, four reactions; Exome target regions listed in Appendix S2). The SeqCap EZ Developer Library was diluted to carry out 16 reactions, as opposed to the commercial protocols suggested four reactions. To do so, PCR water was added directly to the commercial SeqCap EZ bait library bring the total volume to 72 μl (4.5 μl per hybridization). We define hybridization reaction as the process of hybridizing the Nimblegen baits with the DNA to enrich for the chimpanzee exome.

Each pool hybridization reaction was performed by adding 1.5 μg of the equimolar pool of 16 DNA libraries in *Experiment 1* and 0.24 μg of the equimolar pools of 4 DNA libraries in *Experiment 2* (Figure 1), with 5 μl of COT Human DNA (1 mg/ml; contained in the SeqCap EZ Accessory Kit v2) and 2,000 pmol (or 2 μl) of the Multiplex Hybridization Enhancing Oligo Pool (1 μl of 1,000 pmol SeqCap HE Universal Oligo and 1 μl of the 1,000 pmol SeqCap HE Index Oligo pool). The DNA Library Pool/COT Human DNA/Multiplex Hybridization Enhancing Oligo Pool was dried in a DNA vacuum concentrator on high heat (+60°C). Subsequently, we added 7.5 μl of 2X Hybridization Buffer and 3 μl of Hybridization Component A, mixed it by vortexing and heated it to 95°C in a heat block for 10 min to denature the DNA. The Multiplex DNA Sample Library Pool/COT Human DNA/Multiplex Hybridization Enhancing Oligo Pool/Hybridization Cocktail was transferred to a 4.5 μl aliquot of EZ Developer Library (previously diluted) in a 0.2-ml PCR tube, mixed, centrifuged and incubated in a thermocycler at +47°C for 36 hr. Afterwards, we washed following the commercial protocol and amplified the captured DNA using the Post-Capture LM-PCR program, with a total of 12 cycles.

Finally, we performed a second hybridization for the three pool replicates in *Experiment 1* and four pool replicates in *Experiment 2*, as illustrated in Figure 1, following the same protocol as the first hybridization. Only the amount of starting material was altered, using for each of the second hybridizations all the material obtained after the PCR purification from the first hybridization. To limit the extent of PCR duplicates, the captured product of the second hybridization was amplified with eight PCR cycles rather than 12.

2.4 | Sequencing, mapping and on-target reads evaluation

Library pools were merged (as shown in Figure 1) and sequenced in three lanes of an Illumina HiSeq 2500 ultra-high-throughput sequencing system (125-bp paired end); each lane contained one pool from *Experiment 1* and two pools from *Experiment 2*, with each pool contributing a third of the DNA loaded on the lane. For most analyses, the sequencing data were analysed separately by hybridization assay, due to the different conditions carried out in each experiment (i.e., Hyb 1-9 in Figure 1). Adapters from sequenced reads were trimmed using Trim Galore (version 0.4.0) and Cutadapt software (version 1.8.3; Krueger, 2016; Martin, 2011). Reads were aligned to the chimpanzee reference genome panTro4 (Feb. 2011, CSAC Pan_troglodytes-2.1.4 (GCA_000001515.4) using BWA (version 0.7.12) with default alignment parameters (Li & Durbin, 2009). Duplicates were removed after mapping using Picard Tools MarkDuplicates (version 1.95) with default parameters ("http://broadinstitute.github.io/picard/").

To derive high confidence results, we identified what we will hereinafter refer to as "reliable reads." Reliable reads are those that mapped to a single unique genomic location and mapped with a mapping quality score of 30 or higher. Any reference to "mapped reads" will refer to all reads that mapped, reliably or otherwise. A second unique nomenclature is "reliable reads on-target," which are simply reliable reads that mapped to our target space. We obtained the number of reliable reads on-target using the BEDTOOLS INTERSECTBED command (version 2.22.1; Quinlan & Hall, 2010). We intersected the target regions provided by Roche with the reliable reads and then counted the number of reads for each condition using the function samtools -c (SAMTOOLS version 0.1.19; Li et al., 2009). The percentage of reads on-target was calculated by dividing the number of reads on-target by the total number of reads mapped.

The effectiveness of the capture was evaluated by assessing the enrichment factor, capture sensitivity, capture specificity and library complexity. Enrichment factor (EF) was calculated as the ratio of the number of reliable reads on-target and total reads sequenced divided by the ratio between target space (57.5 Mb) and genome size (~3 Gb; Gupta et al., 2010). Capture sensitivity (CS) was defined as the proportion of target regions with an average coverage of at least one, to the total number of target regions; and capture specificity (CSp) was defined as the percentage of unique reads mapping to target sequences, determined by the number of reliable reads on-target divided by the total number of reliable reads (Jones & Good, 2016). Library complexity (LC) was defined as the number of nonduplicated reads divided by the total number of reads mapped, where duplicated reads are those that have identical genomic location on both ends (Chen et al., 2012; Daley & Smith, 2013; Snyder-Mackler et al., 2016).

$$\mathrm{EF} = \frac{\text{Reliable reads on} - \text{target/Total reads}}{\text{Target space (57.5 Mb)/Genome size (3000 Mb)}}$$

$$\mathrm{CSp} = \frac{\text{Reliable reads on} - \text{target}}{\text{Reliable reads}}$$

$$\mathrm{CS} = \frac{\text{Number target regions average coverage} \geq 1}{\text{Total number target regions (295767)}}$$

$$\mathrm{LC} = \frac{\text{Reliable reads}}{\text{Mapped reads}}$$

2.5 | SNP calling, principal component analysis and allele balance

SNPs were called using FREEBAYES (version 0.9.20; Li, 2015) with standard filters and no population priors for each lane of data. Sites with

a quality score below 30 and a depth of coverage (DP) smaller than 4 were removed from further analysis, with the caveat that variants used in the principle component analysis were identified using a less stringent quality score of 20. SNPs were called on both an experimental unit level (N = 72) and on an individual level (N = 17). To call variants for an individual, the BAM files for each experimental unit derived from a unique individual chimpanzee were merged prior to running FREEBAYES. Using VCFTOOLS (version 0.1.12, vcf-isec and vcf-merge; Danecek et al., 2011), we then generated two unique VCF files. One with all experimental library units (N = 72) and a second VCF file combining data for individuals (N = 17) merged with whole-genome sequencing data derived from 59 country-referenced chimpanzees (de Manuel et al., 2016).

The resulting VCF file of genotypes with the combined data from 72 individuals, generated in the previous step, was used in a principal component analysis (PCA) to ascertain the population structure among individuals using PLINK (version 1.90b; Purcell et al., 2007). The VCF file with 72 library units was used in quantifying levels of heterozygosity, genotype distances among individuals and genotype discordance among experimental replicates. Genotyping distances and discordances were estimated from a genotype dosage file, produced by the --012 function in VCFTOOLS. Distances for both dendrogram inference and quantification were estimated by summing the absolute delta of the dosage calls between two libraries and then dividing by the number of markers compared. These latter analyses were carried out using bespoke R scripts.

The VCF file containing the 72 experimental library units was filtered to include only data from on-target regions. From that filtered data set, all heterozygous sites for each individual were identified, and the number of reads that supported the reference allele at each variant as well as the total number of reads was recorded. Finally, the ratio of these two numbers (reference allele observation (RO)/read depth (DP)) was used to evaluate the distribution of allele imbalance, where we would expect an average ratio of 0.5 for balanced data.

2.6 | Technical variation and replicate informativeness

To evaluate the effect that the different variables in *Experiments 1 and 2* have on assay performance, we carried out linear regressions and nested analysis of variance (ANOVA; Fisher, 1936; Gelman, 2005). In each analysis, we used, in turn, the observed number of raw reads acquired, CS, CSp, LC and EF as the response variable and evaluated faeces, extract, library, pool, hybridization, lane, amount of starting material, fragment degradation length and percentage of endogenous content as predictor variables in both univariate and multivariate analysis.

We also carried out subsampling analyses to evaluate library richness (amount of independent, unique and reliable reads) and determine the informativeness of replicates or the level of information gained by employing more than one faecal sample, extract or library from the same individual. BAM file subsampling was carried out in a range from 0.5 M to 6 M, with steps of 0.5 M reads on all experimental libraries (N = 24) in *Experiment 2*. Subsampled read bins of the same size were then merged by extract, and then faeces and finally individuals. As such, 2 M reads for "extract 1" are made up of 1 M reads from both libraries 17 and 18, and similarly, 20 M reads for individual "match1" are composed of 10 M reads from both faeces 1 and faeces 2. The subsampling itself was performed using SAMTOOLS (view-s; version 0.1.19; Li et al., 2009).

3 | RESULTS

3.1 | Sample selection

Across all 48 samples in the screen panel, the estimated mean endogenous content, via the qPCR assay, was 0.78% (median = 0.087%, range = 0.007%–10.74%). Shotgun sequencing data for all samples yielded 7.6 million raw reads, with an average of 160 thousand raw reads per sample. A second estimation of endogenous content is the proportion of total reads that mapped to the reference using the shotgun sequencing data. On average, we observed 2.8% of raw reads mapping to the reference, with a median estimate of 1.1% and a range of 0.16%–24.6%. The proportion of reads that mapped does correlate with the estimated endogenous content of the sample as estimate by qPCR assay (Spearman rho = 0.73, p = 3.04e−9). However, the relationship between these two estimates of endogenous content was better explained by a quadratic function (multiple R^2 = 0.895) than a simple linear regression (multiple R^2 = 0.591) where we forced the intersect through zero (F = 75.31, df = 2, p = 4.36e−15). We chose to move forward with the proportion of shotgun mapped reads as our estimate for endogenous content as we anticipate it being a more accurate estimation of the proportion of molecules in a library that we will be attempting to enrich. The mean of the average fragment lengths, across all samples, was estimated at 1,391 bp (median = 1,356 bp, range = 23–3,700 bp). As such, from the distribution of mean fragment length and endogenous content, as measured by the proportion of mapped reads, we identified 18 faecal samples to move forward with (Appendix S1: Fig. S1). Their mean endogenous content was 4.4% (median = 1.5%, range = 0.19%–24.6%), and their mean average fragment length (degradation) was 1,463 bp (median = 1,408 bp, range = 265–2,452 bp). It is worth noting that no correlation between mean degradation (fragment length) and endogenous content, by either measure, was observed.

3.2 | Sequence data summary

Across the 72 experimental units in this study, we acquired a total of 1,592 million raw reads from three lanes of an Illumina HiSeq 2500; this equates to an average 22.12 million reads, of which 20 million reads were mapped (Figure 2a). However, just 41.4% of the raw reads were duplicate free across all experimental units, where read duplicates are assumed to be the product of PCR amplification during the experiment and thus redundant data.

Of the 9.17 million duplicate free reads, an average of 8.33 million reads was of high quality and deemed "reliable," and 7.40 million or 33.46% of the raw data mapped to our target space. This equates to an average of 66.53% of the acquired data being composed of either PCR duplicates, off-target reads or poor-quality reads.

3.3 | Experiment 1

In *Experiment 1*, we used 16 different faecal DNA extracts from 16 different individuals to create a single pool of 16 DNA libraries for targeted capture sequencing (Figure 1). Each of the three pool replicates was target enriched twice, in parallel. Each of the three hybridized pool replicates was sequenced on a different lane of the HiSeq2500, along with aliquots from *Experiment 2* (Figure 1). Across the 48 (16 × 3) experiments, we acquired an average of 8.3 million reads, of which an average 2.5 million reads (30.2%) were declared reliable, that is, passed quality filters, were duplicate free, and mapped with no secondary alignment. Ninety-four per cent of reliable reads, or 28% of raw reads, were reliable reads that mapped to our target space (Figure 2b). Four other summary statistics that exemplify assay performance are EF, CS, CSp and LC (see Section 2 for nomenclature definitions). Across our 48 assays in *Experiment 1*, we observe an average EF of 12.6, along with a CS of 55%, CSp of 88%, and an average LC value of 0.34.

3.3.1 | Variables affecting performance

Across all *Experiment 1* libraries ($N = 48$), we acquired an average of 8.3 million reads. However, the range of raw data acquired (0.74–45.9 million reads) for each library does vary substantially. In this experiment, each of the 16 libraries were pooled in equimolar ratio, as determined by electrophoretic and flow cytometric assays, followed by targeted hybridization performed in triplicate (Figure 1). As such, under equal conditions, we would anticipate that amount of data acquired from each library would be relatively equal. However, with these NI samples, we found that the count of raw reads acquired for a library was strongly correlated with the endogenous content of the sample (Pearson's $r = .887$, $p = 4.8e{-}17$; Figure 2c). In fact, when we fit the data to a linear regression model forcing the intercept to go through zero, we estimated a Beta value (effect size) of 1.59. That is to say, in this experiment, for each 1% increase in endogenous content we would predict an increase of 1.59 million raw reads acquired (Figure 2c). With the exclusion of "library," all other assay variables, namely sample degradation, total DNA used in library prep, pool and hybridization, each explained significantly less of the overall variation in raw reads acquired in univariate analyses (Figure 3aa). "Library" as an explanatory factor does explain more of the overall variation at 93.7% in raw read counts (Figure 3aa). However, we want to emphasize that endogenous content, average degradation and total DNA used are each

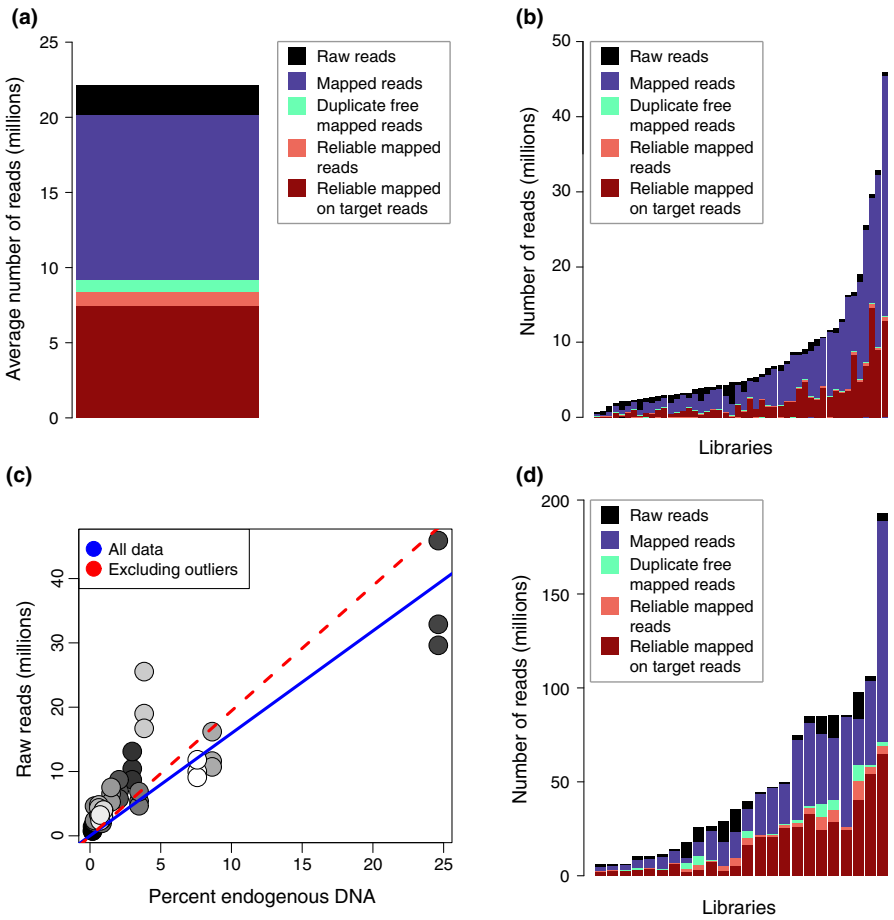

FIGURE 2 Sequencing data summary. (a) Averages across all experiments ($N = 72$) and summaries for each experimental library in (b) Experiment 1 and (d) Experiment 2. (c) Correlation between the percentage of endogenous DNA of each of the 16 libraries from Experiment 1 and the number of raw reads sequenced

summary statistics of, and thus correlated with library. Indeed, we can see in Figure 3ab that when fitting the same data to a multivariate model that only 12.5% of the variation in raw reads is explained by library. In brief, these observations indicates high within replicate similarity and significant distinction among libraries derived from a single biological sample and that sample endogenous content is the single most influential factor in the amount of raw data acquired in complex pools.

Reliable reads correlated with the number of raw reads acquired across all hybridizations in *Experiment 1* (Pearson's $r = .97$, $p = 1.22e{-}21$) and were thus also influenced by endogenous content. However, the proportion of raw reads that were reliable and mapped on-target (mean = 24.2, range = 1.5%–51.3%) were better explained by the hybridization reaction ($\eta^2 = 39.6$, $p = 1.18e{-}5$) than by endogenous content ($\eta^2 = 12.3\%$, $p = 0.0145$). Variation among libraries also had a significant univariate effect on these estimations; however, this signal was attenuated in the multivariate analysis. This observation directly parallels those for EF (Figure 3ac), as the proportion of raw reads that were reliable on-target make up the numerator of the EF calculation, while the denominator is a constant.

The estimation of CS is dependent upon the amount of raw data acquired. As such, we found that the variance explained by each explanatory variable, on CS, correlated with observations for raw reads acquired, as discussed above (Figure 3ae). Yet, in multivariate analysis we observed a much greater effect of some unexplained component of library variation on CS (Figure 3af, $\eta^2 = 59.9\%$, $p = 1.09e{-}27$), with replicates (libraries across hybridizations) correlating very well (Pearson's $r = .99$, $p \ll 1.0e{-}4$). For CSp, we observed a significant univariate effect for endogenous content ($\eta^2 = 0.089$, $p = 0.039$) and total DNA used in library ($\eta^2 = 0.098$, $p = 0.03$), yet it was once again some unexplained component of library variation that accounted for the vast majority of the variation in CSp in both univariate (Figure 3ag; $\eta^2 = 0.95$, $p = 6.13e{-}19$) and multivariate analysis ($\eta^2 = 0.805$, $p = 8.56e{-}18$; Figure 3ah). In contrast, LC is driven by stochastic variation among hybridizations (univariate: $\eta^2 = 0.842$, $p = 8.8e{-}19$; multivariate: $\eta^2 = 0.842$, $p = 3.6e{-}22$; Figure 3ai,j). And yet, LC values correlated among replicates between hybridizations (Pearson's $r > .83$, $p \ll 1e{-}4$) indicating a consistent bias in hybridization performance among libraries within a hybridization experiment.

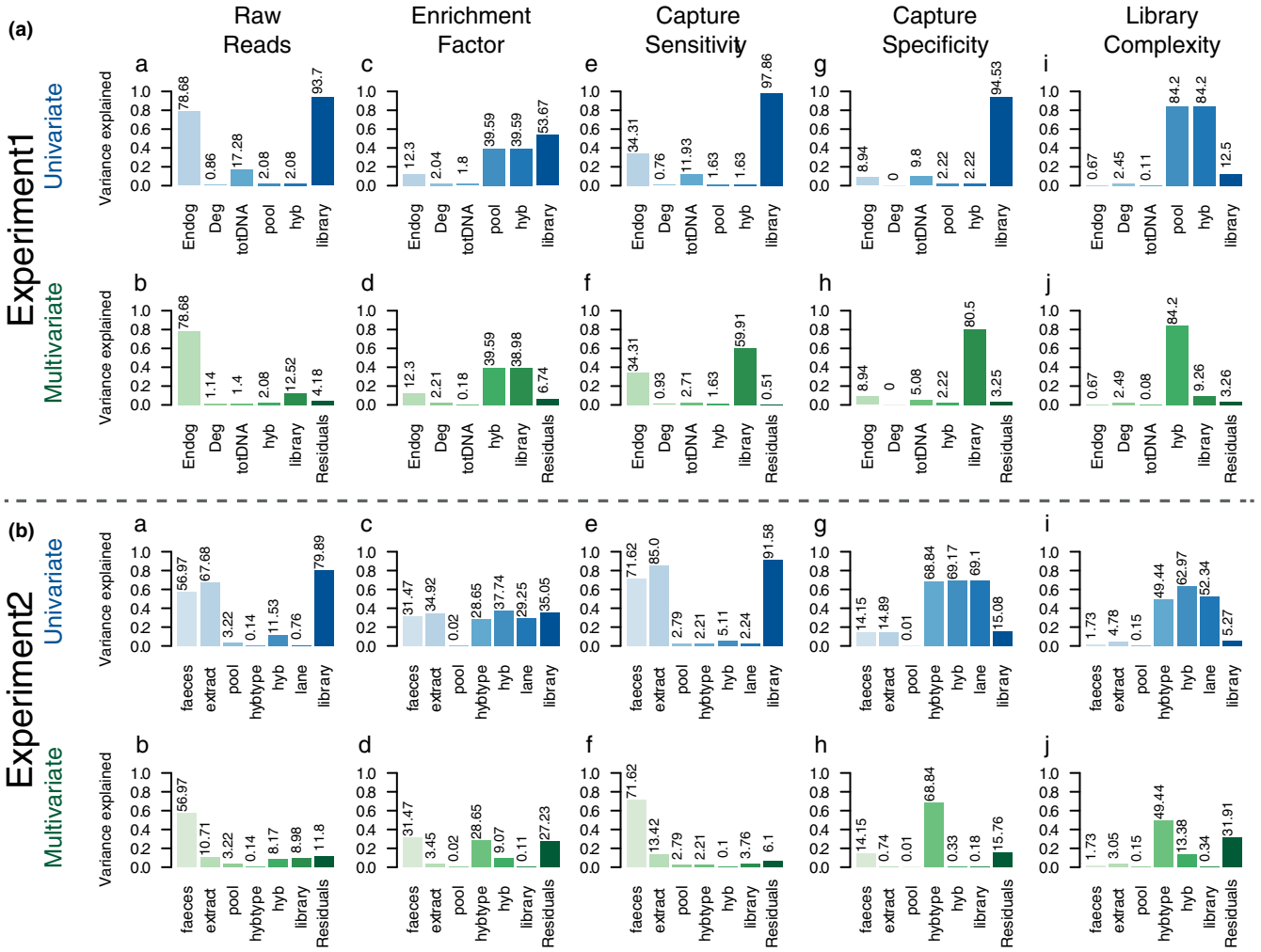

FIGURE 3 Variance explained estimated as the eta-squared statistic from ANOVA sum of squares for univariate and multivariate nested modelling in (a) Experiment 1 and (b) Experiment 2

3.4 | Experiment 2

Experiment 2 was designed to use two samples from a single individual to evaluate the impact that different faecal samples, extracts, libraries, pools and hybridization types had on assay performance and the data acquired. Across the 24 experimental samples in *Experiment 2*, we acquired an average of 49.6 million raw reads (ranging from 6.0 to 192.8 million reads), of which 37.8% were qualified as reliable (range: 17.8%–54.6%), 84.3% of which mapped to our target space (range: 48.5%–95.4%; Figure 2d).

3.4.1 | Variables affecting performance

The variables that appeared most influential across assay performance statistics in *Experiment 2* were the faecal sample used, a proxy for endogenous content, and the hybridization type (one versus two). As seen in *Experiment 1*, it was the proportion of endogenous content that influenced the amount of raw data acquired. Again, we see here that the faecal sample used, a proxy of endogenous content, has a significant influence on the number of raw reads, EF and CS, explaining 56.9% (p = 6.23e−6), 31.5% (p = 2.9e−3) and 71.6% (p = 5.4e−8) of the overall variation across samples in univariate and multivariate modelling (Figure 3ba,c,e and b,d,f), respectively. However, once again, EF, in addition to CSp and LC, is largely influenced by hybridization type, explaining, 28.6% (3.9e−3), 68.8% (p = 1.0e−5) and 49.4% (p = 1.0e−3), of the variation, respectively (Figure 3bc,g,i and d,h,j).

Finally, we examined if performing two rounds of capture is better than performing just one. We have already illustrated that CSp, LC and EF are influenced by the type of hybridization (1 vs. 2). Figure 4 illustrates that two rounds of target enrichment capture increased EF (Figure 4a) and CSp (Figure 4c) but at the cost of LC (Figure 4d). We found that two rounds of capture yielded an average EF of 19.3 and single capture enrichment an average EF of 12.3. These differences are significant when assuming normality of EF (*t* test p = 0.031) and when modelling EF in a univariate generalized linear model with an underlying gamma distribution (chi-square p = 0.008), and as such would suggest that two rounds of capture yield a higher average enrichment factor than one round of capture. However, given that what we are truly after is acquiring enough unique, mapped DNA molecules to call genotypes (discussed below), we further evaluated how two rounds of capture affect the genotype calling. First, we observe that EF correlates with the number of genotyped positions (Pearson's r = .53, p = 0.0076). However, our best predictor of the number of genotypes called was the proportion of our target space covered at our minimum genotype calling depth threshold (Pearson's r = .99, p = 1.4e−27). As such, a better statistic to evaluate whether two rounds of capture are better than one is the number of variable sites genotyped as a function of the number of raw reads acquired (genotype count/number of raw reads). Using this statistic, we found that each raw read yielded on average 0.0259 and 0.0137 genotypes called in the double and single capture experiments, respectively. These values are significantly different assuming a normal distribution (*t* test, p = 0.007) and when modelling the data in a glm with an underlying gamma distribution (chisq, p = .005). As such, we estimate that one would need to acquire ~1.8× as many raw reads in a single capture experiment as in a double capture experiment to genotype the same number of variable positions. Thus, our data indicate that two rounds of target enrichment are more efficient than just one.

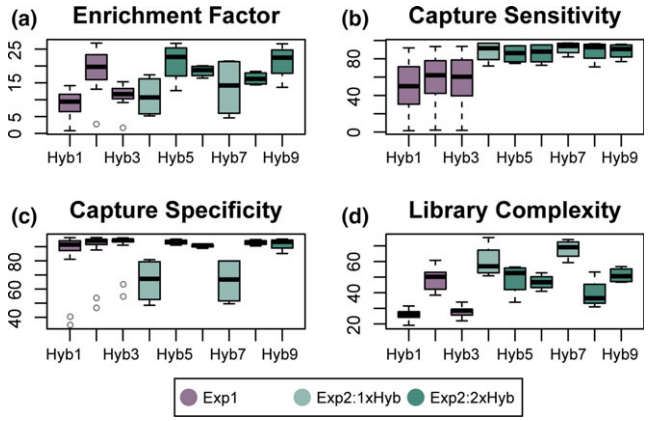

FIGURE 4 Boxplots for (a) enrichment factor or EF, (b) capture sensitivity or CS, (c) capture specificity or CSp and (d) library complexity or LC as grouped by hybridization reaction. The hybridization reaction ids and descriptions are identified in Figure 1

3.4.2 | Library richness

With each sequencing project, library richness is an important aspect of acquiring independent, unique and thus informative base calls for calling variable positions and measuring genetic diversity. In this resampling experiment, we evaluated the extent of library richness or DNA molecule diversity (number of unique sequences that are present in the library). Libraries of low richness will reach an early plateau informing us that deeper sequencing will not provide us with a cost-efficient abundance of unique, independent data. Conversely, libraries that do not reach a plateau retain unique data that can be retrieved by further sequencing. Library richness was determined by subsampling the bam files from 0.5 M to 6 M raw reads (every 0.5 M) for each library from lanes 2 and 3. For all subsampled BAM files, the number of reliable reads was filtered and then combined by library, extract, faecal sample and individual (Figure 5a). We performed this analysis using only the data from *Experiment 2*, hybridizations 5, 6, 8 and 9 (two rounds of capture). We found that the two faecal samples, derived from the same individual, behave quite uniquely (Figure 5a). Faecal sample 2 did not hit a plateau, suggesting that further sequencing on the captured library would continue to provide unique information. In contrast, molecule diversity was approaching exhaustion in the libraries and extracts derived from faecal sample 1. Note that libraries from the two faecal samples were mixed in different pools and hybridized in two hybridization experiments. The depletion of unique molecules with the increase in reads sampled did not correlate with pools or

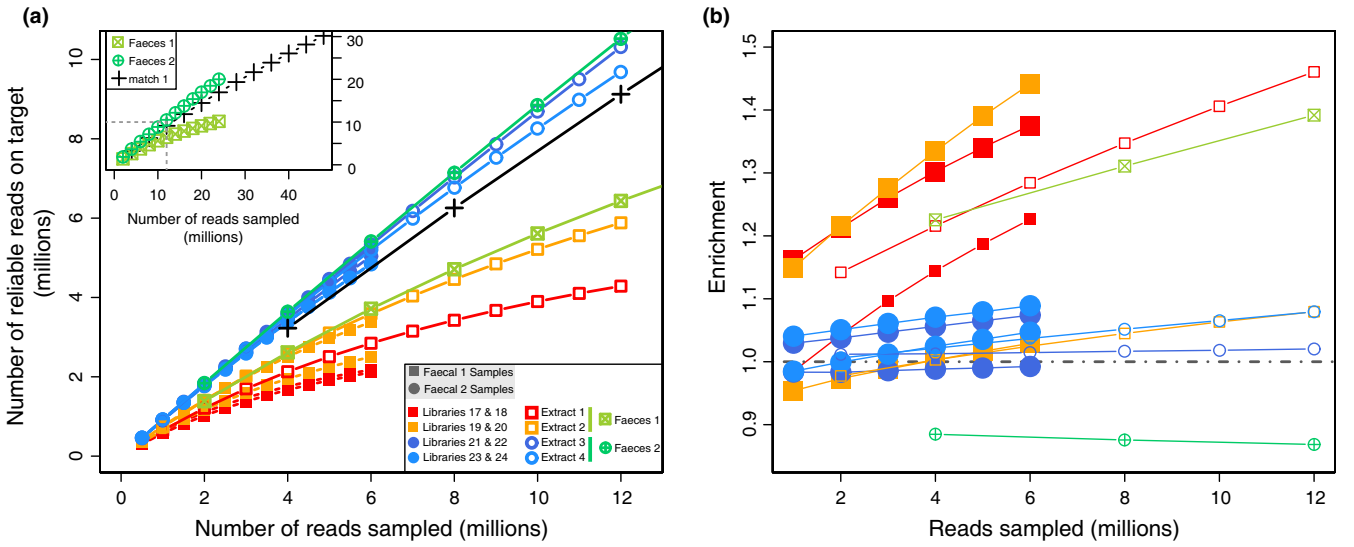

FIGURE 5 (a) Library richness: restricted to samples from Experiment 2; compared are library richness curves for each library, extraction and faeces from the same individual. Plotted are the number of reads subsampled in millions compared to the count of reliable reads mapped to our target space on the y-axis. The inset shows the same plot for the total of 48 million reads subsampled (6 M reads × 8 libraries). (b) Estimations of the relative fold enrichment gained if one sampled a total *N* million reads not from the plotted sample, but *N*/2 million reads from itself and *N*/2 million reads from its replicate pair. Replicate pairs are as follows: (1) Lib17&18, Lib 19&20, Lib21&22, Lib23&24, Extract1&Extract2, Extract3&Extract4, Faeces1&Faeces2

hybridization and was largely the product of the quality of the DNA in the faecal sample. In a complementary analysis, we also calculated the number of unique target regions covered by at least one unique read (Appendix S1: Fig. S3). Interestingly, both faecal samples exhibit similar rates of mapping unique reads to unmapped target regions with increased read sampling. However, faecal sample 2 has a more positive intersect consistent with other observations indicating better sample quality. Regardless, the data suggest that ~10 million raw reads are sufficient data to cover each target base (of the chimpanzee exome) at least once with a uniquely sequenced base.

3.4.3 | Potential gains from using replicates

As we saw above some NI samples may yield capture sequencing libraries that contain limited molecule diversity—a product of their low endogenous DNA content and DNA quality. Yet NI samples are typically hard to come by and precious, and thus, we would seek to maximize the information we may gain from them. As such, we evaluated how much additional information we may gain by processing multiple extracts and/or multiple libraries from a single NI sample (Figure 5b). This evaluation will help us determine the realization of one library or two libraries from the same source when the resources are limited to a certain number of raw reads sequenced. Figure 5b shows that combining libraries 17 and 18 into a single sample yields the extract 1 curve—an overall net increase in the number of reliable reads on-target at any given depth of sequencing. Figure 5b also indicates that a depth of 6 million reads (solid red Libraries 17 & 18), equates to a ~1.3×-fold (30%) increase in information with the addition of the alternative library (if we add library 18 data to library 17) to the experiment. Extracts 1 and 2, both derived from faecal sample 1, were also quite unique – and sure enough if we combined those two libraries in equimolar ratio prior to sequencing we would see a net gain in information (Figure 5b; green faeces 1 curve) at an average of ~1.4× at 12 million raw reads (Figure 5b; average of extracts 1 and 2 curves). Of course, that net gain is relative to the extract sample to start with (Figure 5b). At 12 million raw reads, extract 1 would see a net enrichment of 1.5 with the addition of extract 2 to the experiment. Conversely, given that extract 2 performed better than extract 1 it would see only a 1.1 enrichment with the addition of extract 1 to the experiment. Nevertheless, the only negative enrichment or depletion of information seen is in adding faeces 1 to a study that already contains faeces 2 (Figure 5b, faeces 2 curve). As such, when faced with multiple samples of significantly different quality, these results would suggest the use of better quality samples only. However, in all cases including more libraries from a single extract or even processing multiple extracts from a single NI sample provide more information by increasing overall library richness.

3.5 | Allele imbalance

One major concern with calling alleles with low-quality samples and low coverage data is the relative balance of alleles at a variable position. Here, using allele imbalance estimations across all libraries ($N = 69$), with the exclusion of the N189-10_LR16 contaminated sample, we observed a median estimate of 0.617 indicating a clear bias towards the reference allele (Appendix S1: Fig. S4). This is the same allele used in bait construction. Further, there is significant variation among libraries in their median

estimates of allele balance. An estimated 84.6% of the total variation in median estimates is observed among libraries (*F*-test, $p = 3.75e{-}12$), indicating strong similarities among library hybridization replicates. There is also a clear difference in the allele imbalance distribution between the two faecal samples in *Experiment 2*, with median estimates of 0.65 and 0.57 for faecal samples one and two, respectively (*t* test, $p = 8.8e{-}9$; Appendix S1: Fig. S4B). Overall, observations from both experiments 1 and 2 indicate that allelic imbalance is the product of the biological sample not the hybridization experiment nor the individual. Further, using data strictly from *Experiment 2*, we also observe that two rounds of hybridization does, on average, increase the median allele balance (0.602) as compared to one round of capture (0.589), but not significantly (*t* test, $p = 0.37$). Perhaps most interesting is the observed positive, and nonlinear association between allele imbalance and the number of reliable reads mapped to the target space (spearman rho = 0.80, $p = 2.8e{-}11$) *in Experiment* 1 (Appendix S1: Fig. S4A1), while in contrast there is a negative correlation for these same variables in *Experiment 2* (Pearson's $r = -.715$, $p = 8.46e{-}5$; Appendix S1: Fig. S4B1). But the important contrast between these two analyses is the distribution of reliable reads on-target. When placing all of the data together we observe that up to about 10 million reads on-target, there is a positive association between reads on-target and allele imbalance, but beyond this point the association become negative and perhaps hits a plateau, in our data around 0.56 (Appendix S1: Fig. S4C). We speculate that the positive association is the product of low-depth coverage across sites, which is attenuated, but not balanced, with an increase in coverage.

3.6 | Observed genetic variation

Across both experiments and all samples, we genotyped an average of 914,800 variable sites (range: 300–2.3×10^6) at a minimum call depth of four reads. A single sample, N189-10_LR16, exhibited a unique genotype profile and was excluded for its possible predation–contamination of another primate from the Cercopithecidae family (Watts & Amsler, 2013; Watts & Mitani, 2015; Appendix S1: File S1, Figure S5, S6 and Table S1). Its exclusion reduces the number of variable sites to 394.5 thousand sites on average (range: 138 - 1007x10^6) of which 328.5 thousand sites (range: 119–837.5×10^3) are bi-allelic, on average.

To evaluate the genotype discordance from each of the 72 experiments (24 libraries by 3 lanes) performed in this study, we first removed nine experiments belonging to three samples, GB-18-10_LR1 and GB-36-16_LR9 due to a low number of SNPs called, and N189-10_LR16, because of possible contamination explained above. With the final 63 experiments, we performed a hierarchical cluster analysis and found that each individual library replicate was more similar with itself than with all nonself individuals (Figure 6a). Further, all libraries in *Experiment 2* clustered together and separated by faecal sample. These observations indicate that even with minimum coverage, bias in allele calls and a variable number of markers to compare between libraries, the methodology is robust in replicating the genotype calling of a single biological sample. Importantly though, there are errors in the allele call rates among replicates. On average, library replicates exhibited an average discordance rate of 7.2% with a range from 1% to 9%. In *Experiment* 2, where all libraries are derived from a single chimpanzee, we observed a discordance among libraries within faecal samples to be 9% and 2%, and discordance between faeces to be 6%. Nonself allelic distances, for all libraries in *Experiment* 1, averaged at 21.2%. Indicating that while genotyping errors remain, largely due to the amount of data obtained, distance between different individuals is larger than discordances among replicates.

3.7 | Sample geographic origin

Genetic ancestry was inferred for 14 of 17 individuals by intersecting our data with that of 59 country-referenced chimpanzees (de Manuel et al., 2016). As described above, the two samples with little coverage and the third individual presenting excess variation and Cercopithecidae contamination were not included in this analysis. Recall that study individuals are Central chimpanzee (*Pan troglodytes troglodytes*) from Gabon, and Eastern chimpanzee (*Pan troglodytes schweinfurthii*) from Uganda, while those from de Manuel et al. represent individuals from all four subspecies. Of the 4.6 million sites available from the SNP calling carried out, we kept 65,602 on-target sites that passed quality filters and exhibited <10% genotype missingness. Using PCA, we observe structure largely driven by the uniqueness of Western chimpanzees, as previously described (Gonder et al., 2011; Mitchell et al., 2015). Principle component 2 is driven by variation between central and eastern chimpanzees. As expected, our individuals cluster with their respective subspecies type of Central and Eastern chimpanzees (Figure 6b). This observation is also consistent with our hierarchical clustering analysis that illustrated a monophyletic clade for each subspecies type (Figure 6a).

Data from all samples with the information for statistical analysis and interpretation have been summarized in Table S2.

4 | DISCUSSION

The methodology presented here can be adapted to other designs, such as a selected set of SNPs, whole chromosome, or any targets of choice. However, one must consider the target space of the probes, the quantity of probes per target region and the amount of sequencing that has to be applied to obtain the desirable coverage. For those planning future experiments, our results suggest that endogenous content is the most important factor in this technology. As such, sampling as many specimens as possible will always be ideal. Second, when pooling samples for targeted capturing one should aim to pool samples with the most similar per cent of endogenous content to minimize the drowning of low content

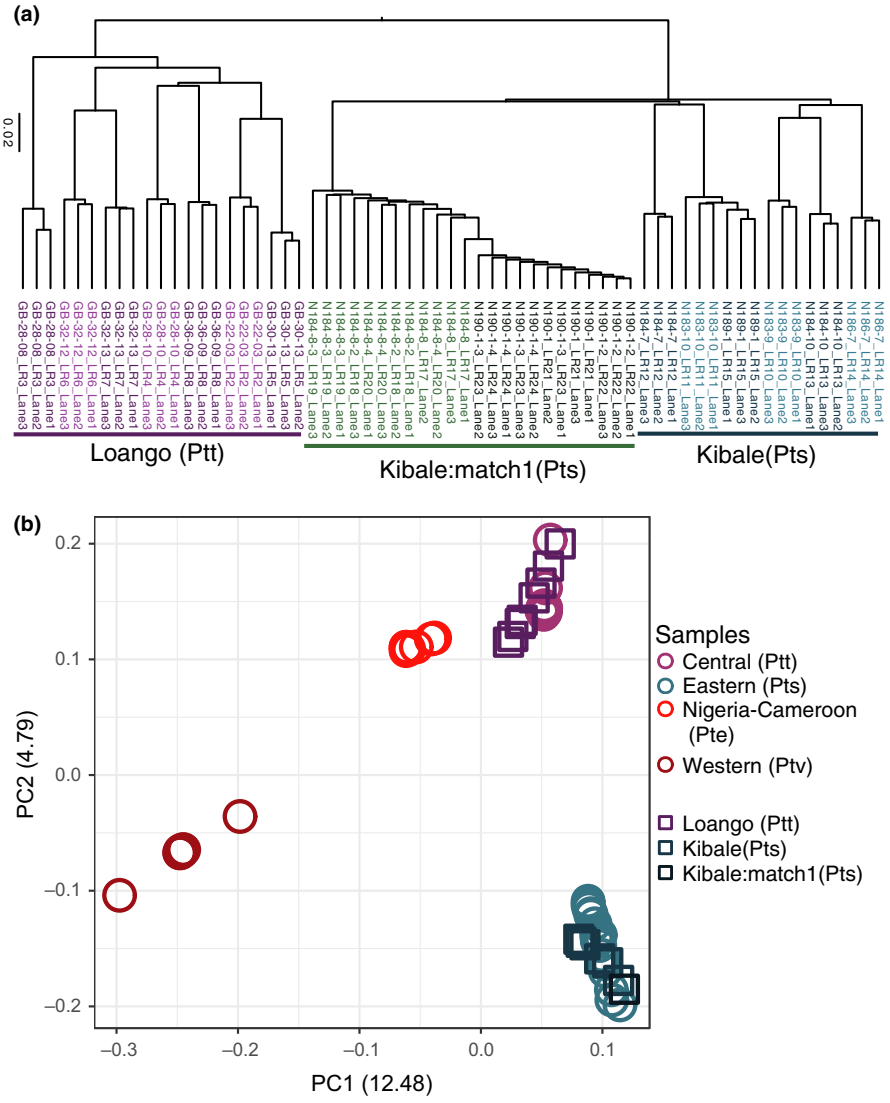

FIGURE 6 (a) Genotype discordance dendrogram from 63 libraries (after removing libraries 1 and 9 due to the low amount of reads and library 16 because of potential contamination). (b) Principal component analysis (PCA) of study samples (squares) along with 59 other chimpanzee individuals from the Great Ape Genome Project (circles)

sample DNA molecules in the pool of those with better endogenous content. Third, when possible perform multiple DNA extracts per specimen and/or create multiple libraries per extract. This will enrich the molecule diversity derived from the sample and as seen in Figure 5a and 5b will help decrease the amount of sequencing necessary for a sample. Note that we are not advising researchers to sequence *N* reads per library, where *N* is the number of reads they estimate needing for a desired coverage. The data are, however, suggesting that sequencing *N*/*i* reads for each "*i*" library derived from a biological sample can pronouncedly increase the yield of unique informative data. This suggestion is additionally strengthened by cost-efficiency of producing extra libraries vs the cost of more sequencing for a single library. Finally, what we are ultimately after is acquiring enough data to reliably genotype our samples. In our data, the number of genotypes correlates with the proportion of target space covered at our minimum calling depth (Pearson's $r = 99.4$, $p = 1.4e{-}6$). As such, to cover ~80% of the exonic target space at depth 4 we require an estimated mean coverage of 20×, and at 40× we cover ~95% of our target space at depth 4 (Figure 7a). These values correspond to ~32 and ~60 million raw reads (Figure 7b). The values are estimates for a target space that is 57.5 Mb. Importantly, the number of raw reads sampled directly influences mean coverage (91.6% of variance explained) and per cent target space covered (55% of variance explained). As such, we reiterate that researchers will gain an advantage by pooling libraries of similar endogenous content or by generating equi-endogenous pools, where the estimates of endogenous content are used to equilibrate among libraries in a pool prior to hybridization and sequencing. Taking such a step will limit variability in data acquired among libraries influenced by endogenous content.

We sequenced and capturing a total of 24 libraries from 17 chimpanzees with a 4-reaction kit targeting the chimpanzee exome, accomplishing two rounds of hybridization from most of the libraries and with replicates. All of the above make our methodology affordable for many laboratories using a commercial kit, without having to produce their own baits. We estimate that the cost from all Roche kits for the library preparation (24 libraries) and hybridization (without sequencing), including clean-up beads and

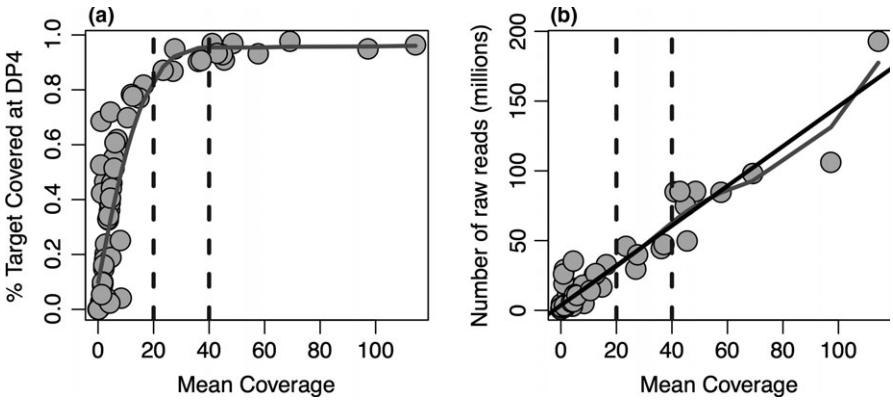

FIGURE 7 (a) Dot plot illustrating the relationship between mean coverage and the percentage of the target space covered at depth 4. Data from all experiments were used ($N = 72$). (b) Correlation between the number of raw reads obtained and the mean coverage in all experiments

purification columns, for the 72 experiments is around 450€ per library.

With our experiments, we have been able to demonstrate that target-capture enrichment can be reliably used to capture target regions from the exome of NI samples. Moreover, we have demonstrated that at least 16 libraries can be pooled and sequenced while still obtaining a considerable number of reads on-target. We have estimated that more genotype data are acquired for less sequencing data when performing two rounds of capture as opposed to one, when assaying NI samples. Moreover, we observe a certain allele imbalance towards the reference allele present in the probes, but we do not discern an elevated difference when comparing between one and two rounds of capture. Further, our data support the production of library replicates to increase data yield as well as the formation of equi-endogenous pools. This latter suggestion will require the development of accurate and robust quantification assays, if not the possibility of low-level shotgun sequencing. Finally, we hope that the evolutionary ecology field at large will find these results and suggestions a utility to their own research.

ACKNOWLEDGMENTS

JH-R is supported by the Ministerio de Economía y Competitividad, Spain (FPI grant BES-2013-064333). This work was supported by the Ministerio de Economía y Competitividad, Spain, and Fondo Europeo de Desarollo Regional (FEDER) (SAF2012-35025 and SAF2015-68472-C2-2-R to FC). The collection of faecal samples was supported by the Max Planck Society Innovation Fund and the Heinz L. Krekeler Foundation's generous funding for the Pan African Programme: The Cultured Chimpanzee. We thank the Agence Nationale des Parcs Nationaux and the Centre National de la Recherche Scientifique (CENAREST) in Gabon and the Uganda National Council for Science and Technology (UNCST) and Ugandan Wildlife Authority (UWA) for their support and permission to collect and export samples from their respective nations. AN is funded by MINECO BFU2015-68649-P (FEDER). TM-B is supported by MINECO BFU2014-55090-P (FEDER), U01 MH106874 grant, Howard Hughes International Early Career, Fundació Zoo de Barcelona and Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya.

DATA ACCESSIBILITY

All sequence data have been submitted to the European Nucleotide Archive and are available under accession code PRJEB21543. https://www.ebi.ac.uk/ena/data/view/PRJEB21543

AUTHOR'S CONTRIBUTIONS

M.A., H.S.K., C.B., L.V., F.C., A.N. and T.M.-B. conceived the study; J.H.-R., D.H. and T.M.-B. designed the study and wrote the manuscript; J.H.-R., M.A., J.L., A.W., M.M., K.L., C.d.F. and S.A. performed experiments; J.H.-R. and D.H. analysed the data sets; J.H.-R., M.A., J.L., C.d.F., A.W., M.M., S.A., F.C., A.N., L.V., H.S.K., K.L., C.B., D.H. and T.M.-B. commented on and approved the final version.

ORCID

Jessica Hernandez-Rodriguez http://orcid.org/0000-0002-0402-286X

REFERENCES

Arandjelovic, M., Guschanski, K., Schubert, G., Harris, T. R., Thalmann, O., Siedel, H., & Vigilant, L. (2009). Two-step multiplex polymerase chain reaction improves the speed and accuracy of genotyping using DNA from noninvasive and museum samples. *Molecular Ecology Resources*, *9*(1), 28–36. https://doi.org/10.1111/j.1755-0998.2008.02387.x

Arandjelovic, M., Head, J., Rabanal, L. I., Schubert, G., Mettke, E., Boesch, C., … Vigilant, L. (2011). Non-invasive genetic monitoring of wild central chimpanzees. *PLoS ONE*, *6*(3), e14761. https://doi.org/10.1371/journal.pone.0014761

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., … Liu, X. S. (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, *9*(6), 609–614. https://doi.org/10.1038/nmeth.1985

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods*, *10*(4), 325–327. https://doi.org/10.1038/nmeth.2375

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFTOOLS. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., … Marques-Bonet, T. (2016). Chimpanzee genomic

diversity reveals ancient admixture with bonobos. *Science (New York, N.Y.)*, *354*(6311), 477–481. https://doi.org/10.1126/science.aag2602

Eriksson, J., Siedel, H., Lukas, D., Kayser, M., Erler, A., Hashimoto, C., … Vigilant, L. (2006). Y-chromosome analysis confirms highly sex-biased dispersal and suggests a low male effective population size in bonobos (*Pan paniscus*). *Molecular Ecology*, *15*(4), 939–949. https://doi.org/10.1111/j.1365-294X.2006.02845.x

Erler, A., Stoneking, M., & Kayser, M. (2004). Development of Y-chromosomal microsatellite markers for nonhuman primates. *Molecular Ecology*, *13*(10), 2921–2930. https://doi.org/10.1111/j.1365-294X.2004.02304.x

Fischer, A., Prüfer, K., Good, J. M., Halbwax, M., Wiebe, V., André, C., … Pääbo, S. (2011). Bonobos fall within the genomic variation of chimpanzees. *PLoS ONE*, *6*(6), e21605. https://doi.org/10.1371/journal.pone.0021605

Fischer, A., Wiebe, V., Pääbo, S., & Przeworski, M. (2004). Evidence for a complex demographic history of chimpanzees. *Molecular Biology and Evolution*, *21*(5), 799–808. https://doi.org/10.1093/molbev/msh083

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Fünfstück, T., Arandjelovic, M., Morgan, D. B., Sanz, C., Breuer, T., Stokes, E. J., … Vigilant, L. (2014). The genetic population structure of wild western lowland gorillas (*Gorilla gorilla gorilla*) living in continuous rain forest. *American Journal of Primatology*, *76*(9), 868–878. https://doi.org/10.1002/ajp.22274

Fünfstück, T., Arandjelovic, M., Morgan, D. B., Sanz, C., Reed, P., Olson, S. H., … Vigilant, L. (2015). The sampling scheme matters: *Pan troglodytes troglodytes* and *P. t schweinfurthii* are characterized by clinal genetic variation rather than a strong subspecies break. *American Journal of Physical Anthropology*, *156*(2), 181–191. https://doi.org/10.1002/ajpa.22638

Gelman, A. (2005). Analysis of variance. *The Annals of Statistics*, *33*(1), 1–53.

Gonder, M. K., Locatelli, S., Ghobrial, L., Mitchell, M. W., Kujawski, J. T., Lankester, F. J., … Tishkoff, S. A. (2011). Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(12), 4766–4771. https://doi.org/10.1073/pnas.1015422108

Gupta, T., Marlow, F. L., Ferriola, D., Mackiewicz, K., Dapprich, J., Monos, D., … Hayat, M. (2010). Microtubule actin crosslinking factor 1 regulates the balbiani body and animal-vegetal polarity of the zebrafish oocyte. *PLoS Genetics*, *6*(8), e1001073. https://doi.org/10.1371/journal.pgen.1001073

Hans, J. B., Haubner, A., Arandjelovic, M., Bergl, R. A., Fünfstück, T., Gray, M., … Vigilant, L. (2015). Characterization of MHC class II B polymorphism in multiple populations of wild gorillas using non-invasive samples and next-generation sequencing. *American Journal of Primatology*, *77*(11), 1193–1206. https://doi.org/10.1002/ajp.22458

Hofreiter, M., Siedel, H., Van Neer, W., & Vigilant, L. (2003). Mitochondrial DNA sequence from an enigmatic gorilla population (*Gorilla gorilla uellensis*). *American Journal of Physical Anthropology*, *121*(4), 361–368. https://doi.org/10.1002/ajpa.10186

Inoue, E., Akomo-Okoue, E. F., Ando, C., Iwata, Y., Judai, M., Fujita, S., … Yamagiwa, J. (2013). Male genetic structure and paternity in western lowland gorillas (*Gorilla gorilla gorilla*). *American Journal of Physical Anthropology*, *151*(4), 583–588. https://doi.org/10.1002/ajpa.22312

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, *25*(1), 185–202. https://doi.org/10.1111/mec.13304

Kanthaswamy, S., Kurushima, J. D., & Smith, D. G. (2006). Inferring Pongo conservation units: A perspective based on microsatellite and mitochondrial DNA analyses. *Primates*, *47*(4), 310–321. https://doi.org/10.1007/s10329-006-0191-y

Krueger, F. (2016). Babraham Bioinformatics - Trim Galore! Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kühl, H. S., Kalan, A. K., Arandjelovic, M., Aubert, F., D'auvergne, L., Goedmakers, A., … Fisher, L. E. (2016). Chimpanzee accumulative stone throwing. *Nature Publishing Group*, https://doi.org/10.1038/srep22219

Langergraber, K. E., Rowney, C., Schubert, G., Crockford, C., Hobaiter, C., Wittig, R., … Vigilant, L. (2014). How old are chimpanzee communities? Time to the most recent common ancestor of the Y-chromosome in highly patrilocal societies. *Journal of Human Evolution*, *69*(1), 1–7. https://doi.org/10.1016/j.jhevol.2013.12.005

Li, H. (2015). FERMIKIT: Assembly-based variant calling for Illumina resequencing data. *Bioinformatics*, *31*(22), 3694–3696. https://doi.org/10.1093/bioinformatics/btv440

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and SAMTOOLS. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lobon, I., Tucci, S., de Manuel, M., Ghirotto, S., Benazzo, A., Prado-Martinez, J., … Marques-Bonet, T. (2016). Demographic history of the genus pan inferred from whole mitochondrial genome reconstructions. *Genome Biology and Evolution*, *8*(6), 2020–2030. https://doi.org/10.1093/gbe/evw124

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Mitchell, M. W., Locatelli, S., Ghobrial, L., Pokempner, A. A., Sesink Clee, P. R., Abwe, E. E., … Gonder, M. K. (2015). The population genetics of wild chimpanzees in Cameroon and Nigeria suggests a positive role for selection in the evolution of chimpanzee subspecies. *BMC Evolutionary Biology*, *15*, 3. https://doi.org/10.1186/s12862-014-0276-y

Morin, P. A., Wallis, J., Moore, J. J., Chakraborty, R., & Woodruff, D. S. (1993). Non-invasive sampling and DNA amplification for paternity exclusion, community structure, and phylogeography in wild chimpanzees. *Primates*, *34*(3), 347–356. https://doi.org/10.1007/BF02382630

Nater, A., Arora, N., Greminger, M. P., van Schaik, C. P., Singleton, I., Wich, S. A., … Krützen, M. (2013). Marked population structure and recent migration in the critically endangered Sumatran orangutan (*Pongo abelii*). *Journal of Heredity*, *104*(1), 2–13. https://doi.org/10.1093/jhered/ess065

Nsubuga, A. M., Robbins, M. M., Roeder, A. D., Morin, P. A., Boesch, C., & Vigilant, L. (2004). Factors affecting the amount of genomic DNA extracted from ape faeces and the identification of an improved sample storage method. *Molecular Ecology*, *13*(7), 2089–2094. https://doi.org/10.1111/j.1365-294X.2004.02207.x

Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., & Hedrick, P. W. (2010). Conservation genetics in transition to conservation genomics. *Trends in Genetics*, *26*(4), 177–187. https://doi.org/10.1016/j.tig.2010.01.001

Perry, G. H., Marioni, J. C., Melsted, P., & Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. *Molecular Ecology*, *19*(24), 5332–5344. https://doi.org/10.1111/j.1365-294X.2010.04888.x

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., … Marques-Bonet, T. (2013). Great ape genetic

diversity and population history. *Nature*, *499*(7459), https://doi.org/10.1038/nature12228

Primmer, C. R. (2009). From conservation genetics to conservation genomics. *Annals of the New York Academy of Sciences*, *1162*(1), 357–368. https://doi.org/10.1111/j.1749-6632.2009.04444.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Quinlan, A. R., & Hall, I. M. (2010). BEDTOOLS: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rogers, J., & Gibbs, R. A. (2014). Comparative primate genomics: Emerging patterns of genome content and dynamics. *Nature Reviews. Genetics*, *15*(5), 347–359. https://doi.org/10.1038/nrg3707

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, *30*(2), 78–87. https://doi.org/10.1016/j.tree.2014.11.009

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., … Tung, J. (2016). Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. *Genetics*, *203*(2), 699–714. https://doi.org/10.1534/genetics.116.187492

Steiner, C. C., Putnam, A. S., Hoeck, P. E. A., & Ryder, O. A. (2013). Conservation genomics of threatened animal species. *Annual Review of Animal Biosciences*, *1*(1), 261–281. https://doi.org/10.1146/annurev-animal-031412-103636

Swenson, J. E., Taberlet, P., & Bellemain, E. (2011). Genetics and conservation of European brown bears *Ursus arctos*. *Mammal Review*, *41*(2), 87–98. https://doi.org/10.1111/j.1365-2907.2010.00179.x

Taberlet, P., Waits, L. P., & Luikart, G. (1999). Noninvasive genetic sampling: Look before you leap. *Trends in Ecology & Evolution*, *14*(8), 323–327. https://doi.org/10.1016/S0169-5347(99)01637-7

Thalmann, O., Fischer, A., Lankester, F., Pääbo, S., & Vigilant, L. (2007). The complex evolutionary history of gorillas: Insights from genomic data. *Molecular Biology and Evolution*, *24*(1), 146–158. https://doi.org/10.1093/molbev/msl160

Thalmann, O., Hebler, J., Poinar, H. N., Pääbo, S., & Vigilant, L. (2004). Unreliable mtDNA data due to nuclear insertions: A cautionary tale from analysis of humans and other great apes. *Molecular Ecology*, *13*(2), 321–335. https://doi.org/10.1046/j.1365-294X.2003.02070.x

Thalmann, O., Serre, D., Hofreiter, M., Lukas, D., Eriksson, J., & Vigilant, L. (2004). Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. *Molecular Ecology*, *14*(1), 179–188. https://doi.org/10.1111/j.1365-294X.2004.02382.x

Vaidyanathan, G. (2011). Apes in Africa: The cultured chimpanzees. *Nature*, *476*(7360), 266–269. https://doi.org/10.1038/476266a

Wall, J. D., Schlebusch, S. A., Alberts, S. C., Cox, L. A., Snyder-Mackler, N., Nevonen, K. A., … Tung, J. (2016). Genomewide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Molecular Ecology*, *25*(14), 3469–3483. https://doi.org/10.1111/mec.13684

Watts, D. P., & Amsler, S. J. (2013). Chimpanzee-red colobus encounter rates show a red colobus population decline associated with predation by chimpanzees at Ngogo. *American Journal of Primatology*, *75*, 927–937. https://doi.org/10.1002/ajp.22157

Watts, D. P., & Mitani, J. C. (2015). Hunting and prey switching by chimpanzees (*Pan troglodytes schweinfurthii*) at Ngogo. *International Journal of Primatology*, *36*(4), 728–748. https://doi.org/10.1007/s10764-015-9851-3

Wultsch, C., Waits, L. P., Hallerman, E. M., & Kelly, M. J. (2015). Optimizing collection methods for noninvasive genetic sampling of neotropical felids. *Wildlife Society Bulletin*, *39*(2), 403–412. https://doi.org/10.1002/wsb.540

Wultsch, C., Waits, L. P., & Kelly, M. J. (2014). Noninvasive individual and species identification of jaguars (*Panthera onca*), pumas (*Puma concolor*) and ocelots (*Leopardus pardalis*) in Belize, Central America using cross-species microsatellites and faecal DNA. *Molecular Ecology Resources*, *14*(6), 1171–1182. https://doi.org/10.1111/1755-0998.12266

Xue, C., Raveendran, M., Harris, R. A., Fawcett, G. L., Liu, X., White, S., … Rogers, J. (2016). The population genomics of rhesus macaques (*Macaca mulatta*) based on whole-genome sequences. *Genome Research*, *26*(12), 1651–1662. https://doi.org/10.1101/gr.204255.116

Yu, N., Jensen-Seaman, M. I., Chemnick, L., Ryder, O., & Li, W.-H. (2004). Nucleotide diversity in gorillas. *Genetics*, *166*(3), 1375–1383. https://doi.org/10.1534/genetics.166.3.1375

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Hernandez-Rodriguez J, Arandjelovic M, Lester J, et al. The impact of endogenous content, replicates and pooling on genome capture from faecal samples. *Mol Ecol Resour*. 2018;18:319–333. https://doi.org/10.1111/1755-0998.12728